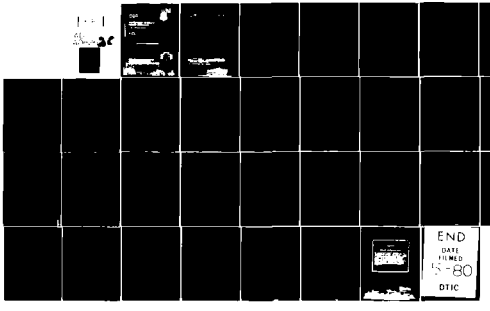
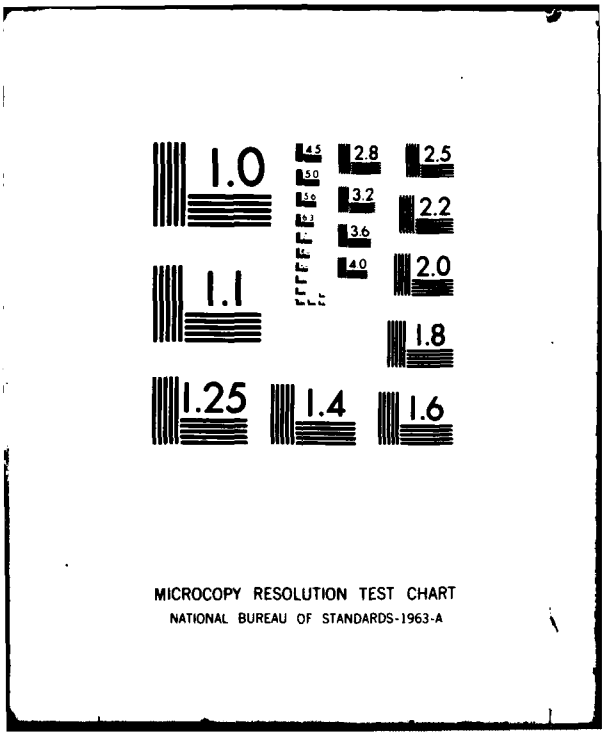
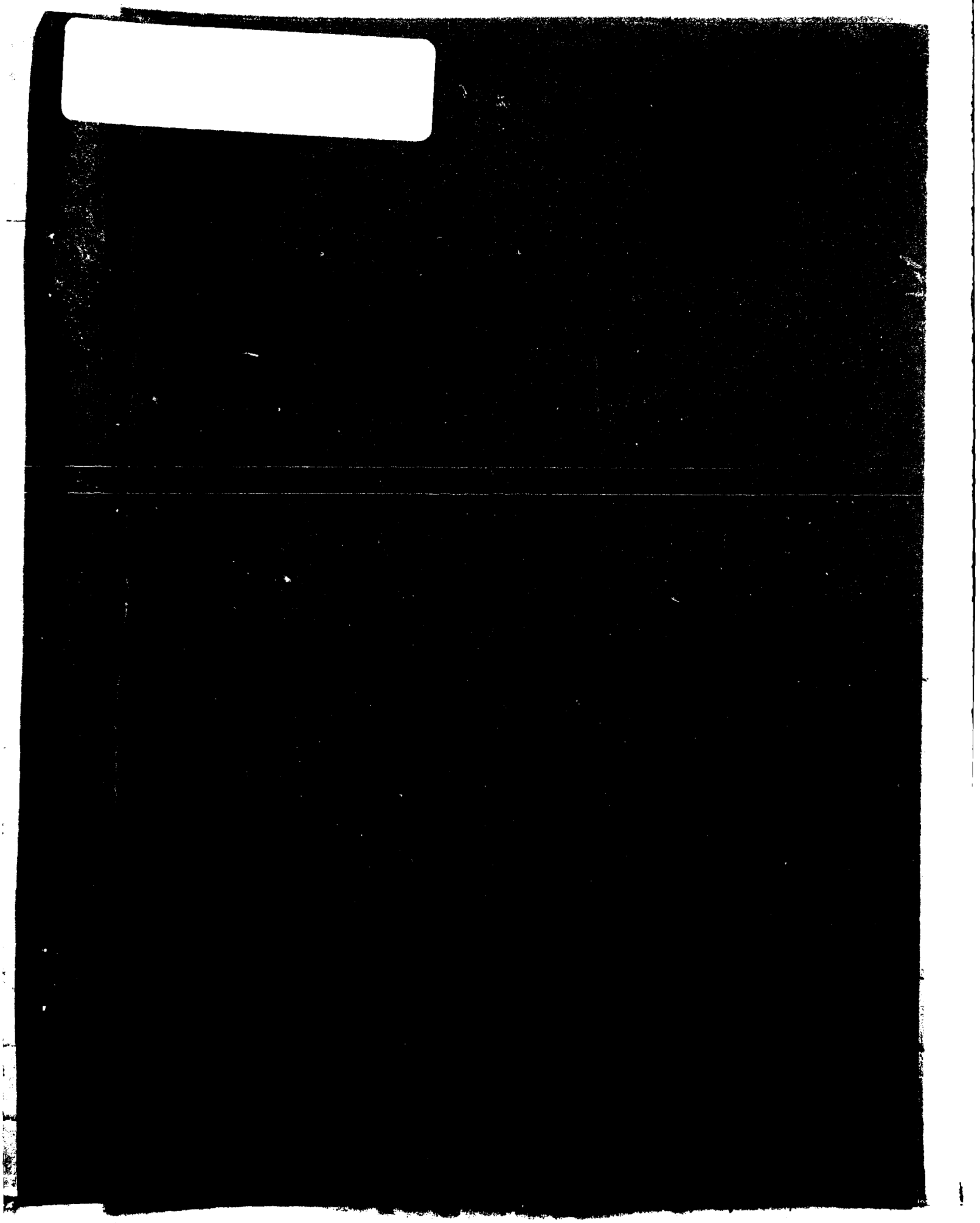
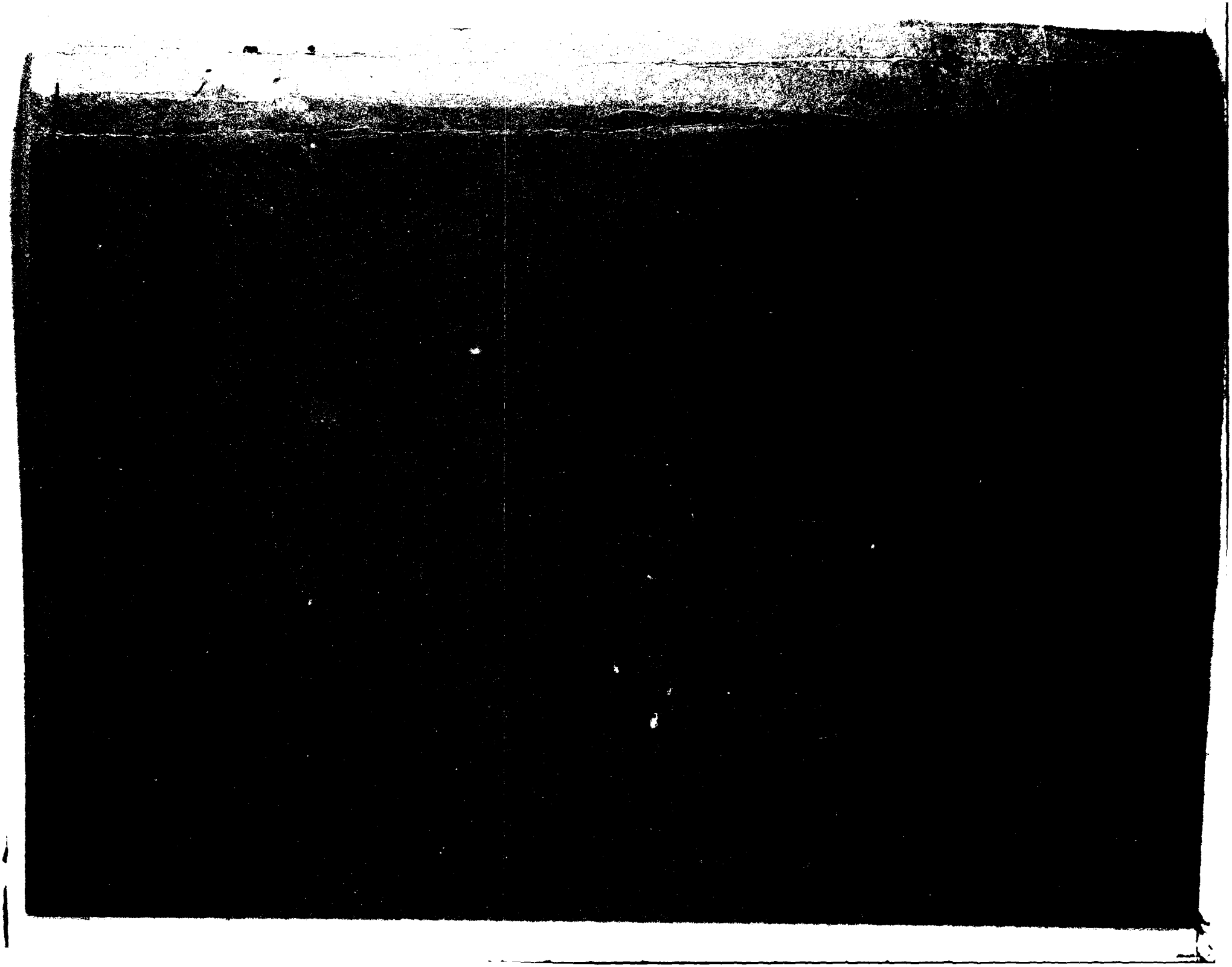
AD-A082 620 PERCEPTION TECHNOLOGY CORP WINCHESTER MA F/G 9/8
CONTINUOUSLY AUTOMATIC SPEAKER ADAPTATION.(U)
JAN 80 L FERBER, H YILMAZ, H KELLETT F30602-78-C-0243
UNCLASSIFIED RADC-TR-79-340 NL



END
DATE FILMED
5-80
DTIC


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
6.3
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A





UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER: RADC-TR-79-349	2. GOVT ACCESSION NO. / 3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle): CONTINUOUSLY AUTOMATIC SPEAKER ADAPTATION	5. TYPE OF REPORT & PERIOD COVERED: Final Technical Report, Jul 1978 — August 1979
	6. PERFORMING ORG. REPORT NUMBER: N/A
7. AUTHOR(s): Leon Ferber, Alex Poohovskoy, Huseyin Yilmaz, Henry Kellett	8. CONTRACT OR GRANT NUMBER(s): F30602-78-C-0243
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Perception Technology Corporation, 95 Cross Street, Winchester MA 01890	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 31011G, 70557998
11. CONTROLLING OFFICE NAME AND ADDRESS: Rome Air Development Center (IRAA), Griffiss AFB NY 13441	12. REPORT DATE: January 1980
	13. NUMBER OF PAGES: 50
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): Same	15. SECURITY CLASS. (of this report): UNCLASSIFIED
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE: N/A

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

Same

18. SUPPLEMENTARY NOTES

RADC Project Engineer: Richard S. Vonusa (IRAA)

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Speech Recognition
Pattern Recognition
Acoustic Phonetics

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

An exploratory development model was implemented and a gisting scenario was devised utilizing a 50-word vocabulary simulating an air traffic control environment. The gisting scenario is an on-line voice-controlled program for the creation, appending and editing of gisting files while an operator is monitoring an air traffic control channel by means of a headset. Recognition results were obtained on the 50-word vocabulary which included ten connected digits plus control words and descriptors simulating the air traffic control (Cont'd)

DD FORM 1 JAN 73 1473

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

Item 20 (Cont'd)

environment. The vocabularies were subdivided in the manner of their use in the gisting scenario with fewer than 24 vocabulary words active at any time. A voice-operated DISABLE/ENABLE facility was provided to prevent the entry of irrelevant conversation into the gisting algorithm. Although real-time recognition was not fully achieved, on-line performance was successfully demonstrated and high recognition accuracy obtained over the entire 50-word vocabulary.

TABLE OF CONTENTS

Section	Title	Page
	INTRODUCTION	2
1.0	EXPLORATORY DEVELOPMENT MODEL	4
	1.1 Recognition Algorithm	5
	1.2 Gisting Scenario	8
2.0	EXPERIMENTAL SET-UP AND PROCEDURES	13
	2.1 Test Phrases	13
	2.2 Data Recording	13
	2.3 Test Procedure	17
3.0	RESULTS	
	3.1 Performance Data	20
	3.2 Analysis of Performance Data	24
4.0	CONCLUSION	28
	REFERENCES	29

TABLE OF FIGURES AND TABLES

Figure 1	Gisting Scenario Block Diagram	10
Table I	Gisting Scenario Vocabularies	14
Table II	Training Utterances	15
Table III	Test Utterances	16
Table 3-1	Results for Connected Digits	21
Table 3-2	Results for The Three Subsets of The Vocabulary	22
Table 3-3	Comparison of 1978 and 1979 SRS	23
Table 3-4	Confusion Matrix for Connected Digits for 46 Male and Female Speakers	25
Table 3-5	Confusion Matrix for Connected Digits for 9 Female Speakers	26
Table 3-6	Confusion Matrix for Connected Digits for 37 Male Speakers	27

EVALUATION

The objective of this program is to develop, test, and evaluate a speaker independent, on-line continuous/isolated speech recognition method for gisting audio/speech material. The system has the capability of recognizing a vocabulary consisting of isolated words as well as connected phrases spoken in English in an unconstrained manner, independent of speaker and of the channel.

Two sets of recognition tests were performed. The first test processed 46 different subjects (males and females) uttering a set of 48 connected triple digit sequences. The testing resulted in 97.5 percent correct digit recognition for the triple digits spoken in a normal manner.

The second test involved 25 different speakers testing a 40 word air traffic control vocabulary. A 99 percent correct word recognition score was achieved for this vocabulary.

This technology shall be used as an aid to analysts in various Air Force Command and Control functions.

Richard S. Vonusa

RICHARD S. VONUSA
Project Engineer

INTRODUCTION

The exploratory development model was implemented on a PDP-11/70 computer and incorporates basic algorithms developed on previous programs. A gisting scenario was developed for the collection of air traffic control data by means of voice input. The recognition vocabulary was increased to 50 words by the addition of control and descriptor words used to control the scenario and enter simulated air traffic control data.

On-line operation has been achieved in a smoothly operating entry procedure, but real-time has not been fully achieved due to the difficulty and time requirement of converting present Fortran programs into faster operating machine language equivalents.

Speaker and channel independence have been improved due to the speaker trainability of the algorithm. This algorithm was deemed to be appropriate for the air traffic control gisting scenario since operators would have an opportunity to train the system to their voices prior to actually utilizing the gisting mode. They could retain their template files for future operating sessions, and could update or remake their files at any time. A new training session would be required if a new operator or a different channel frequency characteristic were to be accommodated.

Substantial improvements have been made in recognition accuracy by broadening the highest frequency channel of the analyzer filter bank, and by eliminating from the recognition process those

quiet sounds which are comparable in spectral magnitude to the background noise. The broader filter now responds more reliably to /s/, /z/, etc. improving overall recognition accuracy. It was found also that previously obtained examples contained irrelevant and inconsistent spectral transitions as the quiet sounds emerged above or dropped below the noise level. A considerable improvement in accuracy was realized by raising the noise threshold and by taking measures to minimize the levels of noise sources, particularly those containing strong spectral peaks.

A gisting scenario has been developed and successfully demonstrated. It is an on-line algorithm for performing a simulated gisting task. In the simulated task, gisting files are created, edited, appended and stored in computer memory. Except for startup and file access, operation is entirely by voice. The task that is simulated is that of gisting in an air traffic control environment. An operator listens to an air traffic control channel by means of a headset, and makes entries by voice of certain types of information. In the simulated task, gisting entries comprise a descriptor word spoken in unconnected form followed by a connected digit group which may be of any length that can be spoken within a 2.5 second time window.

Recognition experiments were performed utilizing recordings of persons speaking randomly selected sets of descriptors followed by digit groups. These were first recorded on audio tape, then transcribed, one word or digit group at a time into computer disk memory for later automatic collection of results.

Template files were constructed from a set of training utterances. These utterances were not used in the tests. Tests were run for each vocabulary using the template files that would be accessed automatically by means of the scenario. Recognition results were tabulated, and overall results compiled for each vocabulary over 46 people. Overall performance was 97.6% on digits in connected groups, and 99% averaged over all command and descriptor words.

1.0 Exploratory Development Model.

An exploratory development model (EDM) was constructed consisting of hardware and software, primarily utilizing the PDP-11/70 computer system. The EDM incorporates basic algorithms developed under previous contracts, and described in previous reports.[1,2]

The hardware configuration is as shown in Figure 1, consisting of the PDP-11/70 CPU with an RP04 40 MBYTE disk and a floating point processor. A CRT terminal is used for instructing the operator through training and recognition modes and an RP04 disk system is used for storage of files including programs, template files, and spectral representations of input utterances. Speech input is via audio recorder or directly from a microphone or telephone line.

Speech is first preemphasized, then passed to inputs of the

parameter extractors which include a 16 channel filter bank plus voice and pitch extractors. These are digitized and converted to computer format in the Data Acquisition System, which is connected to the PDP-11/70 unibus.

The EDM, as described above was implemented on a DEC PDP-11/70 and written in FORTRAN IV+. It is an integral part of an overall system which includes a TU-15, 9 track tape drive, additional CRT and hard copy terminals and a high-speed printer. There is interfaced to the PDP-11/70, a PDP-11/10 dedicated to the control of input and output of a GT-42 interactive graphic terminal.

1.1 Recognition Algorithm.

A speaker trainable algorithm was used since high accuracy was desired in the gisting environment. The basic speaker trainable algorithm developed on an earlier contract is described in previous reports[1,2,3,4]. Changes and improvements were made resulting in faster and more reliable operation. These changes were directed toward the achievement of real-time, on-line operation, speaker and channel independence, and higher accuracy of recognition. Although real-time operation was not realized, the major portion of the above objective was achieved.

1.1.1. Real-Time, On-Line Processing.

During the contract, extensive effort was devoted to the realization of on-line processing, but real-time performance was not realized. Real-time operation necessitates the reprogramming

of all algorithms into time efficient machine language programs, and also the extensive utilization of time sharing between the data input and data processing segments. While simple in concept, reduction to real-time was not realized during the contract due to the extent and complexity of needed programming.

1.1.2. Speaker and Channel Independence.

The nature of the air traffic control gisting task suggested the use of a speaker trainable algorithm, since only a single training session would usually be needed by a new operator of the gisting program. The speaker trainable algorithm has the advantage of high accuracy on people who train it, and also, it obviates the need for speaker and channel transformations, since these are inherent in the training process. It was found, however, that channel and background noise could, if large in amplitude and concentrated into narrow spectral bands, reduce the accuracy of recognition. Further discussion of this is included in the following subsection.

1.1.3. Improvements in Recognition Accuracy.

It was found after carefully examining the results on the previous contract that there was an excessive number of errors involving "six" and "seven", particularly when spoken by women. This was found to be due to a concentration of /s/ energy above 5KHz and practically no energy within the range of the 16 analyzer filters. The /s/ was, in effect, not "heard" by the recognition program and there was too little remaining spectral information for accurate identification of words containing /s/.

This problem was practically solved by simply broadening the 16th filter, thereby making it more responsive to energy above 5KHz. Formerly, this center frequency was 4325 Hz, and it was changed to a broader response with a center frequency at 6000 Hz.

The above change was made prior to making any templates for carrying out the present contract, since templates made with the former filter complement would not properly match the new input utterances. As a result of this change, accuracy has improved overall, even in many utterances not containing /s/.

Inconsistencies were found in the spectral representations used on the previous contract. These seemed to occur in those regions where the raw input amplitude was low, though still of definite significance. The problem was identified as the appearance of erroneous spectral transitions due to the mixture of low amplitude speech and spectrally peaked background noise. A peaked source of noise was identified in the room where data was usually recorded. The problem was practically solved by 1) the removal of all spectrally peaked sources from rooms used in recording, and 2) raising of the noise threshold above the level at which remaining background and channel noise would have significant effect.

Marked improvement in performance has been demonstrated as a result of the above modifications. In many cases, only one digit template is needed for each digit to achieve good recognition accuracy. This was usually not possible in the earlier system.

1.2 Gisting Scenario.

The gisting scenario simulates the task of gisting in an air traffic control environment. The task is for the entry, by voice, of air traffic control information by an operator while listening to an air traffic control channel through a headset. The vocabulary and scenario comprise a representative subset of possible air traffic control tasks. The on-line gisting scenario was developed during the contract and was successfully demonstrated to representatives of the contracting agency.

1.2.1. The Gisting Task.

Gisting files are made and stored in computer memory as a record of air traffic control activity. Entries are made into gisting files by an operator who is listening to an air traffic control channel. Gisting data could be entered via keyboard, but in this case the operator must be a skilled typist. The subject contract was for the creation of a feasibility model for the use of voice input to the gisting files.

1.2.2. Simulation of the Air Traffic Control Environment.

Simulated gisting tasks were devised, and specific modes and vocabulary subsets were defined to accommodate the tasks. Since the resulting system was experimental, additional test modes were incorporated into the scenario. Two gisting modes were incorporated corresponding to two particular types of file data, a) a descriptor word (altitude, time, etc.), followed by a digit group up to 2.5 seconds in length, and 2) an alphabet character (alpha, bravo, etc.)

followed by a digit group. The simulated gisting task comprises operation of the gisting scenario by voice using control words and the entry, by voice, of lists of phrases according to the above two formats. Except for initiation and operator changes, the scenario is controlled entirely by voice without the use of any keyboard functions.

1.2.3. The Gisting Vocabulary.

Vocabulary words used in the scenario are shown in Table I. The digits are recognized in connected mode while all other words have to be spoken singly. Experimental results are given by the use of Control Group 1, while the live demonstration used Control Group 2. In addition to normal gisting, each vocabulary, except Control Group 1, can be accessed for testing within the gisting scenario.

1.2.4. Gisting Scenario Block Diagram.

Figure 1 is a block diagram of the gisting scenario. "Initiation" comprises the keyboard entry of a speaker code such as "HK1" for example. The CRT displays the question "Do I know you" to which a response of "yes" or "No" is entered by voice. If "Yes" is spoken, the operator begins to make normal scenario entries. If "No" is spoken, the training mode is accessed, however, this segment of the scenario was incomplete at the end of the report period.

Normal operation of the scenario is begun in response to the question "What would you like to do?", and a list of options. Voice entry of "New" opens a new gisting file, while entry of "Old"

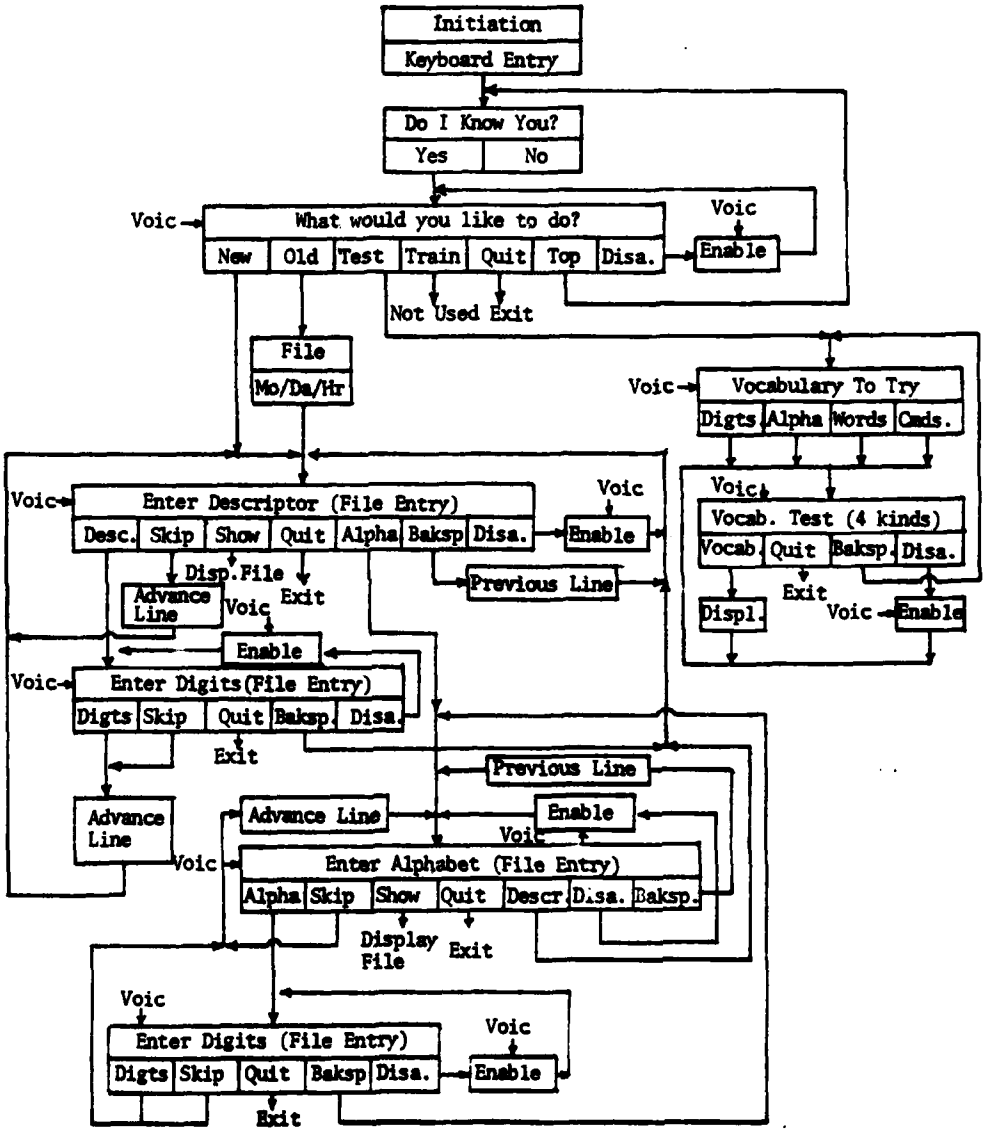


Figure 1: Gisting Scenario Block Diagram.

accesses an existing gisting file. In the latter case, the file identity is entered by keyboard in the form of month/day/hour when the file was created.

The test mode does not create a gisting file, but permits the operator to test the various vocabularies. The scenario instructs him through the operation of accessing each vocabulary and speaking any of the words in the selected vocabulary. Recognition results are displayed on the CRT for visual verification.

The "train", "quit", and "top" functions are obvious from the diagram, although as mentioned above, the automatic training function has not been completed at the time of this report. There is an "disable/enable" operation which inhibits the response and entry of irrelevant speech when such is desired by the operator. Note that this function appears at each entry node, thus permitting the operator to suspend voice entry at any point of the scenario.

The scenario signals its readiness to accept entries into a gisting file by the instruction "Enter descriptor". The operator may enter gisting data or access additional functions either prior to file entries or after any number of entries. The additional functions are "skip", which enters a blank line or terminates a partially complete line of gisting data, "show" which displays the latest form of the gisting file, "quit" which exits the program, "alphabet" which switches to the alphabet descriptor vocabulary, and "backspace", which permits reentry of the previous line. The "disable/enable" functions are, of course operative at this node.

When the "alphabet" command is spoken, control is transferred to the "Enter alphabet" node which is a second descriptor node operating on the same gisting file. Return to the normal descriptor node is by the spoken "Descriptor" command. In other respects, the "Enter alphabet" node is identical to the above described "Enter descriptor" node.

Gisting file entry comprises a descriptor input followed by a digit group, or an alphabet character followed by a digit group. Note that there are identical "Enter digits" nodes depending upon whether a normal descriptor or an alphabet character is first accessed. In addition to digit group entry, additional functions are available at these nodes. "Skip" causes termination of the gisting data line with no digit entry; "Quit" exits the program; "Backspace" deletes the present line and permits its reentry by succeeding commands; and, of course, the "disable/enable" function.

A line of gisting data comprises a descriptor (normal or alphabet) followed by a group of connected digits. Program control is strictly by voice. Upon recognition of a descriptor, control goes to the "Enter digits" node. Upon recognition of a digit group, control advances to the next line via an "Advance line" block for entry of the next line into the gisting file.

This section has described the gisting scenario as developed during the report period. A computer printout of the operational scenario and a description of its operation are given in Appendix A.

2.0 EXPERIMENTAL SET-UP AND PROCEDURES

2.1 Test Phrases.

Recordings were made on magnetic tape of utterances to be used as training and design examples. Training examples consisted of two examples of each vocabulary subset shown in Table I with the addition of a set of digit pairs for use in training the machine to recognize connected digits. Table II is a list of the training utterances as recorded by each test speaker.

Test phrases were randomly chosen by drawing chips from a dish without replacement. Table III is one example. In the case of control words, there are five examples spoken of each control word. It will be noted that each column of the table is in a different random order. Normal mode phrases are made up of pairs of utterances, a descriptor and an appropriate group of connected digits. Alpha phrases are made up of pairs of an alpha character and a group of three connected digits. There are 50 control word utterances, 144 normal mode utterances (72 phrases) and 96 alpha utterances (48 phrases) for a total of 290 test utterances or 170 phrases. A connected digit group was considered to be a single utterance.

Training and test utterances were recorded at the same sitting and under the same conditions. Conditions of the recordings are described in the next subsection.

2.2 Data Recording.

All recordings were made in a quiet area and particular effort was made to assure a reasonable acoustic environment, free of

TABLE I

GISTING SCENARIO VOCABULARIES

Digits	Control Group 1	Control Group 2	
0	yes	old	commands
1	no	new	words
2	entry	test	quit
3	file	train	enable
4	return	top	disable
5	stop	backup	alphabet
6	exit	digits	
7	normal		
8	skip		
9	go		

Descriptor Words		Alphabet Words	
altitude	number	alpha	golf
airspeed	temperature	bravo	hotel
beaconcode	right	Charlie	India
time	left	delta	Juliette
aircraft	forward	echo	kilo
departure	reverse	foxtrot	lima
holding	register		
release	heading		
sector	runway		

PLEASE GIVE YOUR NAME.

PLEASE READ THESE COLUMNS FROM TOP TO BOTTOM. PAUSE AFTER EACH WORD.

0	0	38	38	yes	yes	old	old
1	1	31	31	no	no	new	new
2	2	18	18	entry	entry	test	test
3	3	28	28	file	file	train	train
4	4	48	48	return	return	top	top
5	5	68	68	stop	stop	backup	backup
6	6	88	88	exit	exit	digits	digits
7	7	41	41	normal	normal	commands	commands
8	8	61	61	skip	skip	words	words
9	9			go	go	quit	quit
						enable	enable
						disable	disable

PLEASE READ THESE COLUMNS FROM TOP TO BOTTOM. PAUSE AFTER EACH WORD

altitude	altitude	alpha	alpha
airspeed	airspeed	bravo	bravo
beaconcode	beaconcode	charlie	charlie
time	time	delta	delta
aircraft	aircraft	echo	echo
departure	departure	foxtrot	foxtrot
holding	holding	golf	golf
release	release	hotel	hotel
sector	sector	india	india
number	number	juliette	juliette
temperature	temperature	kilo	kilo
right	right	lima	lima
left	left		
forward	forward		
reverse	reverse		
register	register		
heading	heading		
runway	runway		

Table II. Training Utterances.

A. CONTROL WORDS

Normal	No	Skip	File	Normal
skip	go	file	go	return
go	normal	yes	exit	file
exit	skip	normal	yes	exit
no	entry	no	normal	stop
yes	return	exit	no	skip
stop	exit	entry	stop	entry
return	yes	return	skip	no
entry	stop	go	return	yes
file	file	stop	entry	go

B. NORMAL MODE PHRASES

Departure	Release	Sector	Temperature	Departure	Heading	Reverse	Holding
0425	308	769	83	2024	359	22	8000
airspeed	time	runway	number	temperature	right	time	aircraft
180	1807	29	287	57	51	2139	154
right	holding	heading	register	altitude	release	release	beaconcode
84	7000	155	74	3000	121	971	759
left	temperature	departure	time	holding	left	register	temperature
35	38	0436	1706	6000	80	92	47
aircraft	register	aircraft	reverse	airspeed	forward	heading	runway
738	34	941	76	275	49	163	46
runway	forward	right	forward	reverse	register	sector	right
72	40	17	42	88	44	432	42
sector	reverse	airspeed	release	runway	beaconcode	departure	airspeed
658	48	425	637	89	985	0819	250
heading	beaconcode	beaconcode	holding	number	aircraft	left	number
087	658	595	700	955	521	75	644
altitude	number	left	altitude	time	sector	forward	altitude
4000	588	04	400	0340	731	63	2000

C. ALPHA PHRASES

Foxtrot	Lima	Foxtrot	Hotel	Lima	Juliette	Delta	Golf
700	660	882	633	774	978	365	394
delta	alpha	juliette	india	hotel	charlie	charlie	bravo
552	935	340	559	230	306	159	217
india	charlie	bravo	echo	golf	echo	india	foxtrot
990	691	147	957	025	363	269	495
kilo	echo	lima	alpha	foxtrot	delta	lima	echo
341	961	968	782	105	210	791	611
hotel	juliette	kilo	delta	alpha	kilo	hotel	juliette
408	146	962	654	966	473	330	545
golf	bravo	charlie	golf	bravo	india	kilo	alpha
823	197	521	670	941	183	703	594

Table III. Test Utterances

excessive echoes and reverbration. Some recordings were made in our laboratory, some in private homes, and some in suitable rooms of public buildings.

Recordings were made at 3 3/4 inches per second on one channel of a TEAC 2300S stereo recorder, using an Electro-Voice Model 651 portable microphone system. The operator briefed and coached each test subject, and monitored his recording by means of a headset. If a speaker made an error in reading, or mispronounced a word, the operator stopped the tape recorder and re-recorded the erroneous utterance.

Speakers fit roughly into three categories:

1) Those experienced with Speech Recognition.

2) Those experienced with microphone use.

3) The general public.

There were men and women in all three categories. The best results were obtained for categories 1) and 2), which will be seen from the results of Section 3.

2.3 Test Procedure.

The data base was generated as a set of template files, such that each speaker had one template file for each vocabulary subset (command words, descriptors, alphabet characters, and connected digits). These were used in conjunction with a word definition file for each vocabulary subset, to obtain results indicating the accuracy of the recognition algorithms.

2.3.1. Data Base Generation.

The data base was generated by hand with the aid of a GT-42

Interactive Graphics Terminal. The audio tape recorder output was fed to the recognizer filter bank and its associated parameter circuits. The operator entered one utterance at a time (word or connected digit group) from the tape into the computer. The result was displayed as an amplitude plot and a plot of correlation against the spectrum at peak amplitude. Templates were made by observing the syllabic structure of the utterance, then entering selected samples; selected samples were taken by means of the time-warp computer algorithm corresponding to the first and second half of each syllable. In some cases, there were too few selected samples to allow for breaking the syllable up into two parts, and in that case only a single template was made for the entire syllable. There were, of course, corresponding word files in which word definitions were entered for every valid combination of template sequences.

In the case of connected digits, a number of alternative templates were usually required for some of the digits to achieve the desired accuracy, but for the remaining alphabet subsets, very few alternative templates were used at all. For the most part, the template making task for all alphabet subsets of each person required on the average about an hour's time.

2.3.2. Testing.

Test utterances were transferred from audio tape to the RP04 disk so that automatic testing could be done. Each utterance was carefully stored in its own file under control of the operator. Once the utterances were on the disk, an entire vocabulary subset

could be run automatically, thereby producing the raw test results which are analyzed in Section 3.

The operator brought in the proper template file and word file for each test, corresponding to speaker and vocabulary subset, then ran the automatic recognition program. The recognition program compared each input utterance against each template file entry to obtain a series of syllable responses and their corresponding scores. These were then applied to the word criteria portion of the recognition program for further refinement and ultimately a decision as to the identity of the input utterance.

3.0 RESULTS

3.1 Performance Data.

The results are presented in two sections. The first section presents the raw data as printed out by the speech recognition system. The second section is an analysis of the results by word and speaker categories.

Fifty speakers were recorded as described in the previous section, the results are tabulated in Table 2-1. The table shows results for 46 speakers, the other four recordings were unacceptable because over 25% of the utterances recorded contained pauses. The pauses make the results more indicative of a discreet word recognition system and therefore were not used. The results in Table 3-1 are for the connected digits. There were 48 utterances of 3 digits each for a total of 144 digits for each of the 46 speakers. The overall correct recognition rate including errors from all sources was 97.49%. The results for the upper 90 percentile (4 worst speakers removed) was 98%.

The full vocabulary of command words, Descriptor words and the Alpha words was tested for 25 speakers. These were selected as good talkers based on their performance in the connected digit tests. The results for the three vocabulary subsets are shown in Table 3-2. The data indicates an overall correct recognition of 99% for the whole vocabulary of 40 words. Broken down by subset the results show that the Alpha and Descriptor vocabularies were recognized with 99.1% and the Control vocabulary with 98.9% correct recognition.

Speaker #	Name	Total # of errors	% errors	Speaker #	Name	Total # of errors	% errors
1	JR	8	94.4	24	BK	1	99.3
2	AD	1	99.3	25	DS	0	100.0
3	TS	1	99.3	26	DO	4	97.2
4	WO	1	99.3	27	RH	5	96.5
5	ST	1	99.3	28	JK	1	99.3
6	WR	1	99.3	29	FH	1	99.3
7	LM	1	99.3	30	HK	0	100.0
8	LC	4	97.2	31	JV	4	97.2
9	JB	6	95.8	32	AK	4	97.2
10	EM	4	97.2	33	TN	1	99.3
11	BL	5	96.5	34	MB	0	100.0
12	JM	4	97.2	35	CH	5	96.5
13	MO	0	100.0	36	JR	7	95.1
14	JN	1	99.3	37	FC	3	97.9
15	DC	9	93.8	38	RB	5	96.5
16	DW	6	95.8	39	GA	2	98.6
17	TB	6	95.8	40	DH	8	94.4
18	LF	1	99.3	41	JF	5	96.5
19	BG	3	97.9	42	DM	2	98.6
20	FK	2	98.6	43	AS	7	95.1
21	RP	0	100.0	44	JF	15	89.6
22	OC	5	96.5	45	CW	14	90.3
23	FM	3	97.9	46	RC	0	100.0

Table 3-1: Results For Connected Digits.
Table showing total number of errors and
% correct recognition for each of 46 speakers.

3.1.1. Performance on Old Data.

The objective of this test was to compare the performance of the 1978 and the 1979 systems and to study the effect of speaker variability with time. The data for this test consisted of digits in groups of 1, 2, and 3 digit strings. The results obtained in 1978 were based on a set of templates obtained at the same time as the test data. The 1979 results are based on the 1978 test data

but the training set was obtained in 1979. A summary of the comparison is shown in Table 3-3.

3.1.2. Final Demonstration.

The results of the final demonstration were obtained in two

Speaker #	Speaker Name	Alpha	Command	Descriptors	Average % of correct recognition
1	AD	0	1	1	98.8
2	AK	0	1	2	98.2
3	CW	1	2	1	97.6
4	RC	0	0	1	99.4
5	HK	0	0	0	100.0
6	LF	1	1	1	98.2
7	JK	0	0	0	100.0
8	FK	0	1	1	98.8
9	MB	0	0	0	100.0
10	AS	2	0	1	98.2
11	TS	0	0	1	99.4
12	WO	0	0	0	100.0
13	ST	1	1	0	98.8
14	WR	0	0	1	99.4
15	LM	0	1	1	98.8
16	RP	0	0	1	99.4
17	BK	0	2	0	98.8
18	DS	1	0	0	99.4
19	DO	0	1	0	99.4
20	FH	1	1	0	98.8
21	JV	0	0	2	98.8
22	TN	2	0	1	98.2
23	LC	0	2	1	98.2
24	EM	0	0	1	99.4
25	BG	2	0	0	98.8

Table 3-2: Results For The Three Subsets Of The Vocabulary. Each error represents 1.4% in the Descriptor subset and 2% in the Alpha and Command subsets. The average % correct recognition is for the vocabulary as a whole averaged over the three subsets.

Speaker		% Correct Recognition	
#	Name	1978 SRS	1979 SRS
1	MB	96.9	98.1
2	LF	98.8	98.8
3	AD	98.8	100.0
4	HK	95.6	98.1
5	DD	97.5	97.5
6	OC	95.6	97.5

Table 3-3: Comparison of 1978 and 1979 SRS. (Speech Recognition Systems). Results are based on test data recorded in 1978.

parts, a) spot check of test results, and b) live testing.

For the first part the RADC representative chose two speakers at random from the file. The test results were verified by a test run using the procedure outlined in Section 2.3.2.

The live test consisted of a demonstration of the Gisting operation. The RADC representative trained the system by recording one set of digits spoken in a discreet manner. He also recorded a set of ten double digits sequences. Based on these recordings a set of reference patterns was obtained containing one sample for each digit except the digits 2, 3 and 8 which had two representative patterns each. The test consisted of reading a list of fifty triple digit sequences with a performance accuracy of 93.3%. Most of the errors were in the digits 2 and 8. The digit 8 accounted for 70% of the errors all of them omission type errors, indicating that it had

an insufficient number of representative templates. Two of PTC's personnel, AD and HK, used the system to enter 25 data lines into a table using voice gisting only. AD was able to enter the 25 data items using 27 statements, HK did it using 26 statements.

3.2 Analysis of Performance Data.

The results indicate that an overall accuracy of 97.5% for connected digits was achieved. In order to analyze these results the speakers were categorized into groups according to sex and experienced or nonexperienced talkers.

The confusion matrix for the digits for the whole group of 46 speakers is shown in Table 3-4. The digit "two" accounts for the largets number of omission errors (7.2%). This is due to the fact that a high acceptance threshold was set for this word in order to avoid extraneous recognitions. The extraneous recognitions for the digit "two" were only 1.8% indicating that there is room for a better setting of the threshold. For the digit "eight" this threshold seems optimal as the balance between omission and extraneous errors is 5.4% to 6.6%. The major source of errors was due to the confusion of the digits "one" and "nine", between them they accounted for 13.2% of the errors. The reason for the errors is due to the coarticulation problem that occurs when "one" is preceded by digits ending with a nasal, and "nine" is preceded by digits ending with a u. It is interesting to note that the traditional "five" "nine" confusion was eliminated. This type of error, namely for the digits "one" - "nine", is worst in the case of female speakers. Table 3-5

RECOGNIZED

SPOKEN	0	1	2	3	4	5	6	7	8	9	?
0		1.2	3.0	0.6	3.6			3.0			1.2
1					2.4					6.0	3.6
2	1.2			2.4	0.6						7.2
3			6.6						4.8		
4	1.2	3.6									3.0
5					1.8						0.6
6			0.6	0.6				1.8			1.2
7	2.4	3.0			0.6						
8			2.4	2.4							5.4
9		7.2		0.6							1.2
E		1.2	1.8	0.6	2.4			0.6	6.6	1.2	

Table 3-4: Confusion Matrix for Connected Digits for 46 Male and Female Speakers. The errors are in percent (%), computed by normalizing the number of errors in each element by the total number of errors. The E in the "SPOKEN" column means an extraneous word was recognized even though it was not spoken.

shows the confusion matrix for the 9 female speakers. The overall performance for female speakers was 96.7% versus 97.8% for male speakers however, the distribution of errors was markedly different. Table 3-6 shows the confusion matrix for 37 male speakers. The highest number of errors is due to omission errors by the digit "two" (9.8%) and "three" - "two" confusion (7.3%). For female speakers some of the errors are also due to "two" - "three" confusion (8.9%).

However, the largest number of errors was due to the "nine" - "one" (15.6%) confusion and the "one" - "nine" (13.3%) confusion. The fact that these two digits contributed to 29% of the errors indicates a deficiency in the recognition of nasals.

The results for the Alpha, Descriptor and Control vocabularies indicate correct recognition of 99%. A confusion matrix is not needed since 96% of all errors are errors of omission.

SPOKEN \ RECOGNIZED	0	1	2	3	4	5	6	7	8	9	?
0		4.4		2.2	4.4						
1					4.4					13.3	
2	2.2			8.9	2.2						
3			4.4						6.7		
4	4.4										4.4
5					2.2						
6				2.2				2.2			2.2
7		4.4									
8											2.2
9		15.6									
E									4.4	2.2	

Table 3-5: Confusion Matrix for Connected Digits for 9 Female Speakers. The errors are in percent (%), computed by normalizing the number of errors in each element by the total number of errors. The E in the "SPOKEN" column means an extraneous word was recognized even though it was not spoken.

RECOGNIZED

SPOKEN	0	1	2	3	4	5	6	7	8	9	?
0			4.1		3.3			4.1			1.6
1					1.6					3.3	4.9
2											9.8
3			7.3						4.1		
4		4.9									2.4
5					1.6						0.8
6			0.8					1.6			0.8
7	3.3	2.4			0.8						
8			3.3	3.3							6.5
9		4.1		0.8							1.6
E		1.6	2.4	1.6	3.3			1.6	7.3	0.8	

Table 3-6: Confusion Matrix for Connected Digits for 37 Male Speakers. The errors are in percent (%), computed by normalizing the number of errors in each element by the total number of errors. The E in the "SPOKEN" column means an extraneous word was recognized even though it was not spoken.

The group of experienced talkers chosen from police, fire department personnel and our own laboratory had the best performance record. The 25 speakers shown in this category had an overall correct recognition of 98.4% for connected digits, and 99% for the rest of the vocabulary.

4.0 CONCLUSIONS

A voice activated interactive command and control system is constructed and demonstrated. The system is used as a test bed to evaluate various voice activated command and control tasks as well as performance levels for several groups of speakers. The results are as follows:

1. Connected digits are recognized with an accuracy of 98.4% for talkers with experience in speaking situations.

2. Command and control vocabularies are recognized with an accuracy of 99%.

3. The performance for female speakers is within 1% of the performance for male speakers although the system is optimized for male speakers.

4. For experienced speakers performance results on connected digits are practically the same on old and new test data using one set of templates.

It is not obvious whether this performance level is adequate for many applications but it is felt that the high degree of interaction naturalness and the fast throughput achieved makes this system a good starting point and a test bed for voice command and control applications.

REFERENCES

1. Yilmaz, H. et al, "Speaker Adaptation Test and Evaluation", Final Report, RADC-TR-79-122, May 1979, A018442.

2. Yilmaz, H. et al, "Automatic Speaker Adaptation", Final Report, RADC-TR-76-273, September 1976, A032592.

3. Ferber, L. et al, "Speech Perception", Final Report, RADC-TR-75-265, November 1975, A071675.

4. Yilmaz, H. et al, "Perceptual Continuous Speech Recognition", Final Report RADC-TR-74-180, July 1974 (AD783899).

5. Yilmaz, H. et al, "Speech Perception Research", Final Report, Contract No. DAAB03-72-C-0407, March 1973.

6. Yilmaz, H. et al, "A Real Time, Small-Vocabulary, Connected-Word Speech Recognition System", Final Report RADC-TR-72-281, November 1972 (AD753176).

7. Yilmaz, H., "A Theory of Speech Perception II", Bulleting of Mathematical Biophysics, Vol. 30, 1968.

8. Yilmaz, H., "A Theory of Speech Perception", Bulletin of Mathematical Biophysics, Vol. 29, 1967.

APPENDIX A

Gisting Scenario Operation

This appendix describes the gisting scenario operation with reference to a printout of the responses that would normally appear on the CRT terminal. A test run was made through all of the gisting nodes, and the resulting printout is given on pages A5 through A12.

The gisting program was started by means of keyboard entries, including initialization to speaker HK1. Printed responses to the initialization instructions are shown at the top of page A5. After initialization the program asks, "Do I know you?" to which the operator answers "yes" by voice input. The program then asks "What would you like to do?" and lists the options. The operator answers "new", indicating that a new gisting file is to be created. The normal gisting mode is entered automatically; the program asks for a descriptor; and the operator responds by saying, in this case, "altitude" into the microphone. A descriptor must be followed by a digit group in this mode, and the operator enters the digit group "300", in connected speech. The program displays the gisting file entry "Altitude = 3000" and asks for the next descriptor. Instead of a descriptor, the operator says "quit", causing the program to return to the previous menu and ask again "What would you like to do?" The operator says "disable", but the program does not recognize the utterance. It gives error messages (bottom of page A5) and again prints the main menu at the top of page A6.

The operator says "disable" a second time. The program recognizes it and inhibits further input. When he is ready to resume, the operator speaks "enable", the program is enabled, and the previous menu is again displayed. The operator next says "old" indicating that he wants to append or edit an existing file. Note that the program asks for a typed entry here. The operator types "9, 25, 10" which are the month, day, and hour that the above new file was created. The program asks for a descriptor, but instead the operator says "show" to display the contents of the accessed file. Note here that "altitude 3000" is printed corresponding to the known content of the previous file. The program asks for a descriptor and the operator says "aircraft". The operator then says the connected digit group "279".

The line "Aircraft = 279" is displayed at the top of page A7 then the program asks for the next descriptor. The operator says "alphabet", a command word that switches to the alternate gisting mode (alphabet character followed by digit group). The computer responds with "enter alphabet". The operator says "alpha" which is recognized, then "123", after which the recognized gisting line "alpha = 123" is displayed. The command word "show" is then spoken and the program displays the contents of the gisting file. The first entry was made upon creation of the file, the second and third were added in the "old" file mode, and the third was from the alternative gisting mode using alphabet characters as descriptors. The operator next says "quit" thereby returning to the gisting menu,

then says "test" to enable the test menu which indicates the vocabularies that can be tested.

At the top of page A8, the program asks "which vocabulary do you want to try?" and lists the vocabularies that can be accessed. In this case the operator says "digits", thereby accessing the connected digits mode. The operator speaks six digit groups, illustrating the flexibility of the connected digit recognizer to recognize groups up to 6 or more digits in length. The operator then says "backup", to return to the vocabulary test menu. The program asks "Do you really want to backup?", to which the operator answers "yes". At first the program does not understand what was said, but responds when "yes" is repeated a second time.

The vocabulary test menu appears again at the top of page A9, and the "alphabet" vocabulary is tested. The operator speaks "alpha", "bravo", and "charlie" and the program responds with the correct recognition each time. He then says "backup" and returns to test the vocabulary of "words" or descriptors.

On page A10, the word vocabulary is tested. The operator at first disables the gister by voice, then resumes by speaking "enable". The spoken word input "altitude" is not recognized at first, but is recognized when spoken a second time. Vocabulary words "altitude", "airspeed", "beaconcode" and "time" are spoken and the program responds correctly. Similarly, at the bottom of page A10 the command word vocabulary is accessed for testing.

The command word vocabulary test is done on page A11. All

words in command group No. 2 are spoken and correctly recognized. Certain command words are needed for program control, and when one of these is recognized the program asks: "Do you really want to _____" where "train", "backup" or "quit" are inserted in the blank. This test continues at the top of page A12.

At the end of the command word test at the top of page A12, the operator says "backup" once to get to the vocabulary test menu, then again to get back to the main menu. At this time he says "quit", which must be verified by saying "yes". At this time, the program exits, and the exercise of the gisting scenario is completed.

```
 RUN GISTER
CHANGE NPR,NSH ? (Y,..) >
CHANGE BOUNDARY/THRESH/INTRO  ADJUSTMENT? >
DEBUG CORREL? >
HELLO, THIS IS THE  P.T.C.  GISTER.
PLEASE TYPE YOUR (3) INITIALS.
HK1


DO I KNOW YOU?


WHAT WOULD YOU LIKE TO DO?

NEW     (CREATE NEW FILE)
OLD     (CONTINUE OLD FILE)
TEST    (TEST TEMPLATES)
TRAIN
QUIT
TOP


YOU SAID:::: NEW
             186
  2 >  ENTER DESCRIPTOR  (AIRSPEED,AIRCRAFT,ETC)
             COMMAND     (QUIT,SKIP,BACKUP,SHOW,ALPHABET)


YOU SAID:::: ALTITUD
             180
  2 > ALTITUD =>


YOU SAID:::: 3        0        0
            167      179      185
  2 > ALTITUD = 3          0        0
  3 >  ENTER DESCRIPTOR  (AIRSPEED,AIRCRAFT,ETC)
             COMMAND     (QUIT,SKIP,BACKUP,SHOW,ALPHABET)


YOU SAID:::: QUIT
             186


WHAT WOULD YOU LIKE TO DO?

NEW     (CREATE NEW FILE)
OLD     (CONTINUE OLD FILE)
TEST    (TEST TEMPLATES)
TRAIN
QUIT
TOP
```

```
SORRY,I DIDN'T UNDERSTAND WHAT YOU SAID.

AUDMAT:COULD YOU REPEAT THAT PLEASE

WHAT WOULD YOU LIKE TO DO?

NEW      (CREATE NEW FILE)
OLD      (CONTINUE OLD FILE)
TEST     (TEST TEMPLATES)
TRAIN
QUIT
TOP

YOU SAID:::: DISABLE
             182
THE AUDOMAT IS NOW DISABLED
TO CONTINUE, PLEASE SAY  "ENABLE"
THE GISTER IS ENABLED

WHAT WOULD YOU LIKE TO DO?

NEW      (CREATE NEW FILE)
OLD      (CONTINUE OLD FILE)
TEST     (TEST TEMPLATES)
TRAIN
QUIT
TOP

YOU SAID:::: OLD
             174
TYPE:  ENTER MONTH,DAY,HOUR(0-24),DESIRED FILE 9,25,10
  3 >  ENTER DESCRIPTOR  (AIRSPEED,AIRCRAFT,ETC)
             COMMAND     (QUIT,SKIP,BACKUP,SHOW,ALPHABET)

YOU SAID:::: SHOW
             182
  2: ALTITUD 3         0         0
  3 >  ENTER DESCRIPTOR  (AIRSPEED,AIRCRAFT,ETC)
             COMMAND     (QUIT,SKIP,BACKUP,SHOW,ALPHABET)

YOU SAID:::: AIRCRAF
             183
  3 > AIRCRAF =>
```

```
YOU SAID:::: 2        7        9
             172      182      180
  3 > AIRCRAF = 2         7         9
  4 >  ENTER DESCRIPTOR  (AIRSPEED,AIRCRAFT,ETC)
             COMMAND     (QUIT,SKIP,BACKUP,SHOW,ALPHABET)


YOU SAID:::: ALPHABT
             180
  4 >  ENTER ALPHABET    (ALPHA,BRAVO,CHARLIE,...)
             COMMAND     (QUIT,SKIP,BACKUP,SHOW,DESCRIPTOR)


YOU SAID:::: ALPHA
             184
  4 > ALPHA    =>


YOU SAID:::: 1        2        3
             188      174      180
  4 > ALPHA    = 1         2         3
  5 >  ENTER ALPHABET    (ALPHA,BRAVO,CHARLIE,...)
             COMMAND     (QUIT,SKIP,BACKUP,SHOW,DESCRIPTOR)


YOU SAID:::: SHOW
             190
  2: ALTITUD 3         0         0
  3: AIRCRAF 2         7         9
  4: ALPHA   1         2         3
  5 >  ENTER ALPHABET    (ALPHA,BRAVO,CHARLIE,...)
             COMMAND     (QUIT,SKIP,BACKUP,SHOW,DESCRIPTOR)


YOU SAID:::: QUIT
             186


WHAT WOULD YOU LIKE TO DO?

NEW      (CREATE NEW FILE)
OLD      (CONTINUE OLD FILE)
TEST     (TEST TEMPLATES)
TRAIN
QUIT
TOP


YOU SAID:::: TEST
             185
```

```
WHICH VOCABULARY DO YOU WANT TO TRY?

DIGITS
ALPHABET
WORDS
COMMANDS


YOU SAID:::: DIGITS
             174


PLEASE SAY ANY SUBSET OF THE DIGITS
YOU MAY CONTINUE SPEAKING AFTER I HAVE RESPONDED

IF I MAKE TOO MANY MISTAKES , PLEASE SAY 'TRAIN'   AND RE-TRAIN ME.
OF COURSE YOU CAN SAY 'BACKUP' AND  ASK FOR ANOTHER VOCABULARY


YOU SAID:::: 1
             187


YOU SAID:::: 1        2
             186      186


YOU SAID:::: 1        2        3
             186      174      171


YOU SAID:::: 1        2        3        4
             187      174      177      177


YOU SAID:::: 1        2        3        4        5
             185      177      174      180      185


YOU SAID:::: 1        2        3        4        5        6
             187      174      175      175      185      193


YOU SAID:::: BACKUP
             188


AUDMAT: DO YOU REALLY WANT TO BACKUP  ?


PLEASE REPEAT
```

```
WHICH VOCABULARY DO YOU WANT TO TRY?

DIGITS
ALPHABET
WORDS
COMMANDS


YOU SAID:::: ALPHABT
             186


PLEASE SAY ANY SUBSET OF THE ALPHABT
YOU MAY CONTINUE SPEAKING AFTER I HAVE RESPONDED

IF I MAKE TOO MANY MISTAKES , PLEASE SAY "TRAIN"   AND RE-TRAIN ME.
OF COURSE YOU CAN SAY "BACKUP" AND  ASK FOR ANOTHER VOCABULARY


YOU SAID:::: ALPHA
             188


YOU SAID:::: BRAVO
             190


YOU SAID:::: CHARLIE
             190


YOU SAID:::: BACKUP
             180


AUDMAT: DO YOU REALLY WANT TO BACKUP  ?


WHICH VOCABULARY DO YOU WANT TO TRY?

DIGITS
ALPHABET
WORDS
COMMANDS


YOU SAID:::: WORDS
             184
```

```
PLEASE SAY ANY SUBSET OF THE WORDS
YOU MAY CONTINUE SPEAKING AFTER I HAVE RESPONDED

IF I MAKE TOO MANY MISTAKES , PLEASE SAY "TRAIN"   AND RE-TRAIN ME.
OF COURSE YOU CAN SAY "BACKUP" AND  ASK FOR ANOTHER VOCABULARY


YOU SAID:::: DISABLE
             184
THE AUDOMAT IS NOW DISABLED
TO CONTINUE, PLEASE SAY  "ENABLE"
THE GISTER IS ENABLED



SORRY,I DIDN'T UNDERSTAND WHAT YOU SAID.


YOU SAID:::: ALTITUD
             179


YOU SAID:::: AIRSPEE
             167


YOU SAID:::: BEACONC
             179


YOU SAID:::: TIME
             182


YOU SAID:::: BACKUP
             183


AUDMAT: DO YOU REALLY WANT TO BACKUP  ?


PLEASE REPEAT


WHICH VOCABULARY DO YOU WANT TO TRY?

DIGITS
ALPHABET
WORDS
COMMANDS


YOU SAID:::: COMMANS
             178
```

```
PLEASE SAY ANY SUBSET OF THE COMMANS
YOU MAY CONTINUE SPEAKING AFTER I HAVE RESPONDED

IF I MAKE TOO MANY MISTAKES , PLEASE SAY 'TRAIN'   AND RE-TRAIN ME.
OF COURSE YOU CAN SAY 'BACKUP' AND  ASK FOR ANOTHER VOCABULARY



YOU SAID:::: OLD
             177


YOU SAID:::: NEW
             182


YOU SAID:::: TEST
             186


YOU SAID:::: TRAIN
             191


AUDMAT: DO YOU REALLY WANT TO TRAIN   ?


YOU SAID:::: TOP
             190


YOU SAID:::: BACKUP
             187


AUDMAT: DO YOU REALLY WANT TO BACKUP  ?


YOU SAID:::: DIGITS
             176


YOU SAID:::: COMMANS
             184


YOU SAID:::: WORDS
             186


YOU SAID:::: QUIT
             188
```

```
AUDMAT: DO YOU REALLY WANT TO QUIT    ?


YOU SAID:::: DISABLE
             182
THE AUDOMAT IS NOW DISABLED
TO CONTINUE, PLEASE SAY  "ENABLE"
THE GISTER IS ENABLED


YOU SAID:::: ALPHABT
             184


YOU SAID:::: BACKUP
             186


AUDMAT: DO YOU REALLY WANT TO BACKUP  ?


WHICH VOCABULARY DO YOU WANT TO TRY?

DIGITS
ALPHABET
WORDS
COMMANDS


YOU SAID:::: BACKUP
             177


WHAT WOULD YOU LIKE TO DO?

NEW      (CREATE NEW FILE)
OLD      (CONTINUE OLD FILE)
TEST     (TEST TEMPLATES)
TRAIN
QUIT
TOP


YOU SAID:::: QUIT
             188


ARE YOU SURE YOU WANT TO QUIT?


NICE CHATTING WITH YOU HK1
BYE
TTO  --  STOP
```